SMOOTHED QUANTILE REGRESSION FOR STATISTICAL DOWNSCALING OF EXTREME EVENTS IN CLIMATE MODELING

ZUBIN ABRAHAM*, FAN XIN**, AND PANG-NING TAN*

Abstract. Statistical downscaling is commonly used in climate modeling to obtain high-resolution spatial projections of future climate scenarios from the coarse-resolution outputs projected by global climate models. Unfortunately, most of the statistical downscaling approaches using standard regression methods tend to emphasize projecting the conditional mean of the data while paying scant attention to the extreme values that are rare in occurrence yet critical for climate impact assessment and adaptation studies. This paper presents a statistical downscaling framework that focuses on the accurate projection of future extreme values by estimating directly the conditional quantiles of the response variable. We also extend the proposed framework to a semi-supervised learning setting and demonstrate its efficacy in terms of inferring the magnitude, frequency, and timing of climate extreme events. The proposed approach outperformed baseline statistical downscaling approaches in 85% of the 37 stations evaluated, in terms of the accuracy of the magnitude projected for extreme data points.

1. Introduction

An integral part of climate modeling is downscaling, which seeks to project future scenarios of the local climate based on the coarse resolution outputs produced by global climate models (GCMs). Two of the more common approaches to downscaling are dynamic downscaling and statistical downscaling. Dynamic downscaling uses a numerical meteorological model to simulate the physical dynamics of the local climate while utilizing the climate projections from GCMs as initial boundary conditions. Though it captures the geographic details of a region unresolved by GCMs, the simulation is computationally demanding while its spatial resolution remains too coarse for many climate impact assessment studies. Statistical downscaling establishes the mathematical relationship between the coarse-scale GCM outputs and the fine-scale local climate variables based on observation data. Unlike dynamic downscaling, it is flexible enough to incorporate any predictor variable and is relatively inexpensive. Most of the statistical downscaling approaches employ regression methods such as multiple linear regression, ridge regression, and neural networks to estimate the conditional mean of the future climate conditions. These methods are ill-suited for predicting extreme values of the climate variables.

An alternative approach is to use techniques such as quantile regression, which aims to minimize an asymmetrically weighted sum of absolute errors, to estimate the particular quantile that corresponds to extreme values [27]. Unfortunately, quantile regression tends to overestimate the response variable resulting in a large number of data points being falsely predicted to be extreme. Figure 1 represent the histogram of the distribution of observed temperature at a weather station in Canada. The lines represent the distribution of the predicted values for temperature obtained using multiple linear regression (MLR) and quantile regression. An observation is considered an extreme data point if its response variable is in the top 5 percentile of observations. The shape of the tail of the distribution that represents extreme data points (observed and projected) is shown in Figure 2. It is clear from the figures that methods such as multiple linear regression (green line) that estimate the conditional mean tend to underestimate the tail of observed probability distribution, while quantile linear regression (red line) overestimates the tail part of the probability distribution. As elaborated

*Michigan State University, Dept of Computer Science, abraha84@msu.edu, ptan@cse.msu.edu
**Michigan State University, Dept of Statistic, fanxin@msu.edu.

in Section 5, it was found that for the 37 stations evaluated, at an average, quantile regression predicted a datapoint to be an extreme point more than twice as frequently as the actual frequency of observed extreme data points.

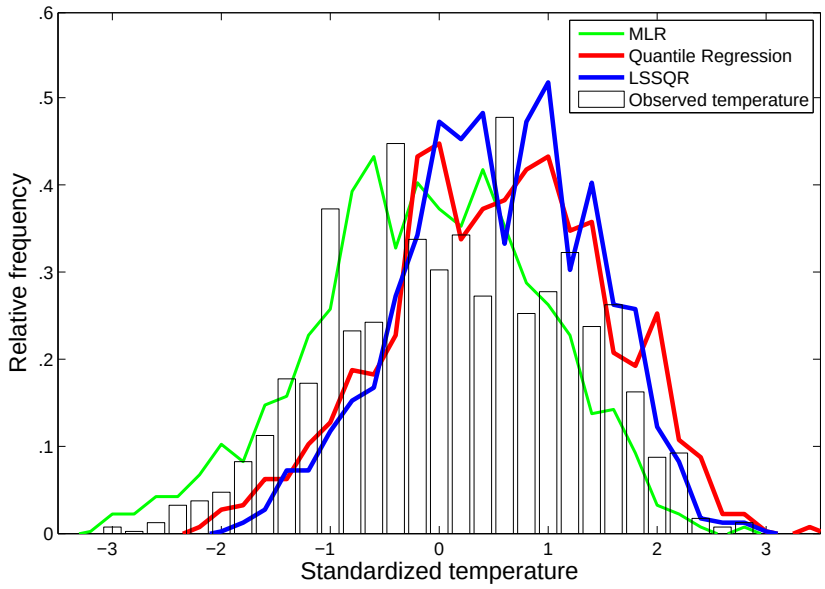

FIGURE 1. Histogram of observed temperature.

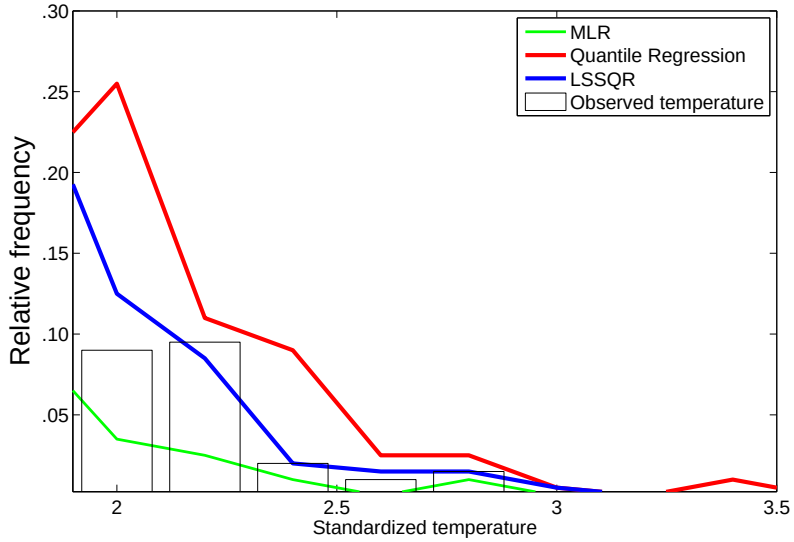

FIGURE 2. Tail of the histogram.

To address this overestimation, we propose a method known as smoothed quantile regression (LSQR) that reduces the absolute error of extreme data points by introducing a smoothing term

that brings the predicted response value of extreme points closer to the value corresponding to the percentile of extreme data points. This smoothing term also provides a means to easily extend the objective function to a semi-supervised learning setting (LSSQR). Semi-supervised learning, in addition to using the training data, can also use the distribution characteristics of the predictor variables of the test set to glean a better estimate of the distribution of data upon which the model will be applied.

In summary, the main contributions of this paper are as follows:

- We demonstrate the limitation of MLR, ridge regression and quantile regression in predicting values for extreme data points.
- We present a smoothed quantile regression framework for extreme values prediction.
- We also extend the framework to a semi-supervised setting.
- We demonstrate the efficacy of our learning framework on climate data (temperature) obtained from the Canadian Climate Change Scenarios Network website [1]. Both the supervised and the semi-supervised proposed frameworks outperformed the baseline methods in 85% of the 37 stations evaluated, in terms of magnitude, frequency and the timing of the extreme events.

The remainder of this paper is organized as follows. Section 2 covers some of the related work. Section 3 introduces the reader to the notations and terminology used in the paper. Relevant approaches, such as quantile regression are also introduced. Section 4 introduces the objective function of the proposed supervised and semi-supervised model, as well as the analysis of the model. This is followed by a detailed description of our algorithm and experimental results in Section 5. Finally, we present our conclusions and suggestions for future work in Section 6.

2. Related Work

Time series prediction has long been an active area of research with applications in finance [41], climate modeling [20][13], network monitoring [11], transportation planning [25], etc. There are several time series prediction techniques available, including least square regression [28], recurrent neural networks [24], Hidden Markov Model Regression [23], and support vector regression [34].

Given the growth in the number of climate models in the earth science domain, extensive research has been done to best utilize these models [32] as well as focus on downscaling surface climate variables like temperature and precipitation time series from these global climate models (GCM) [13, 14, 20, 40]. Identifying and modeling extreme events in climatology has recently gained a lot of traction, especially with regard to temperature [8]. Unfortunately, the common regression techniques mentioned earlier that may be used for downscaling, focus on predicting the conditional mean of the response variable while extreme values are better identified by conditional quantiles as against to condition means. Hence, unlike the common regression techniques mentioned earlier that focus on predicting the conditional mean, the motivation behind the presented model is focusing on a particular conditional quantile, similar to quantile regression [7], so as to accurately predict extreme values during downscaling. Like many of the previous technique mentioned, [7] does not predict the timing of the extreme values.

Variations of quantile regression such as non-parametric quantile regression and quantile regression forests have been used to infer the conditional distribution of the response variable which may be used to build prediction intervals [35, 31]. Also, variants of quantile regression that estimate the median are used due to its robustness to outliers when compared to traditional mean estimate [42]. [22] presented a statistical downscaling approach to estimate censored conditional quantiles of precipitation that uses QR. The conditional probability of the censored variable is estimated using a generalized linear model (GLM) with a logit function to model the nature of the distribution of precipitation and hence cannot be directly applied to model temperature. Mannshardt-Shamseldin et. al. demonstrate another approach to downscaling extremes through the development of a family

of regression relationships between the 100 year return value (extremes) of climate modeled precipitation(NCEP and CCSM) and station-observed precipitation values [29]. Generalized extreme value theory based approaches have also be applied to model extreme events like hydrologic and water quality extremes, precipitation, etc [37, 6]. The Pareto distribution [48, 49], Gumbel [50, 51] and Weibull [52] are the more common variants of General extreme value distribution used. But these techniques are probabilistic based that emphasize trends pertaining to the distribution of future extreme events and not the deterministic timing of the occurrence of the extreme event.

The drawback of building a model that primarily focuses on only a particular section of the conditional distribution of the response variable is the limited amount of available data. Hence, the motivation for incorporating unlabeled data during model building. There have been extensive studies on the effect of incorporating unlabeled data to supervised classification problems, including those based on generative models[19], transductive SVM [26], co-training [9], self-training [45] and graph-based methods [5][46]. Some studies concluded that significant improvements in classification performance can be achieved when unlabeled examples are used, while others have indicated otherwise [9, 16, 18, 36, 43]. Blum and Mitchell [9] and Cozman et al. [16] suggested that unlabeled data can help to reduce variance of the estimator as long as the modeling assumptions match the ground truth data. Otherwise, unlabeled data may either improve or degrade the classification performance, depending on the complexity of the classifier compared to the training set size [18]. Tian et al. [36] showed the ill effects of using different distributions of labeled and unlabeled data on semi-supervised learning.

3. Preliminaries

Let $D_l = \{(x_i, y_i)\}_{i=1}^{n}$ be a labeled dataset of size n, where each $x_i \in \mathcal{R}^d$ is a vector of predictor variables and $y_i \in \mathcal{R}$ the corresponding response variable. Similarly, $D_u = \{(x_i, y_i)\}_{i=n+1}^{n+m}$ corresponds to the unlabeled dataset. The objective of regression is to learn a target function $f(x, \beta)$ that best estimates the response variable y. β is the parameter vector of the target function. n represents the number of labeled training points and m represents the number of unlabeled testing points.

3.1. Multiple linear regression (MLR) and ridge regression.

One of most widely used forms of regression is multiple linear regression. It solves a linear model of the form

$$y = x^T\beta + \boldsymbol{\epsilon}$$

where, $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2)$ is an i.i.d Gaussian error term with variance σ^2. $\boldsymbol{\beta} \in \mathcal{R}^d$ is the parameter vector. MLR minimizes the sum of squared residuals

$$(y - X\beta)^T(y - X\beta)$$

which leads to a closed-form expression for the solution

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^Ty$$

A variant of MLR, called ridge regression or Tikhonov regularization is often used to mitigate overfitting. Ridge regression also provides a formulation to overcome the hurdle of a singular covariance matrix X^TX that MLR might be faced with during optimization. Unlike the loss function of MLR the loss function for ridge regression is

$$(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta,$$

and its corresponding closed-form expression for the solution is

$$\hat{\boldsymbol{\beta}} = (X^TX + \lambda I)^{-1}X^Ty$$

where, the ridge coefficient $\lambda > 0$ results in a non-singular matrix $X^TX + \lambda I$ always being invertible. The problem with both MLR and ridge regression is that they try to model the conditional mean, which is not best suited for predicting extremes.

3.2. **Quantile Linear Regression(QR).** The τ^{th} quantile of a random variable Y is given by:

$$Q_Y(\tau) = F^{-1}(\tau) = \inf\{y : F_Y(y) \geq \tau\}$$

where,

$$F_Y(y) = P(Y \leq y)$$

is the distribution function of a real valued random variable Y and $\tau \in [0, 1]$.

Unlike MLR that estimates the conditional mean, quantile regression estimates the quantile (e.g., median) of Y.To estimate the τ^{th} conditional quantile $Q_{Y|X}(\tau)$, quantile regression minimizes an asymmetrically weighted sum of absolute errors. To be more specific, the loss function for quantile linear regression is:

$$\sum_{i=1}^{N} \rho_\tau(y_i - x_i^T \beta)$$

where,

$$\rho_\tau(u) = \begin{cases} \tau u & u > 0 \\ (\tau - 1)u & u \leq 0 \end{cases}$$

Unlike MLR and ridge regression that have a closed-formed solution, quantile regression is often solved using optimization methods like linear programming. Linear programming is used to solve the loss function by converting the problem to the following form.

$$\min_{u,v} \quad \tau 1_n^T u + (1 - \tau) 1_n^T v$$

$$\text{s.t.} \quad y - x^T \beta = u - v$$

where, $u_i \geq 0$ and $v_i \geq 0$. But as shown in Figures 1 and 2, quantile regression often overestimates data points resulting in too many false positive extreme events predicted.

4. Framework for smoothed quantile regression

Given that the primary objective of the model is to accurately regress extreme valued data points and quantile regression has been shown to perform relatively better that its least square counterparts that tend to underestimate the frequency and magnitude of extreme data points, the proposed objective approach of the proposed frameworks is modeled around linear quantile regression. Section 4.1 describes smoothed quantile regression (LSQR) and its objective function. Section 4.2 proposes a semi-supervised extension to LSQR which is then followed by mathematical properties of the behavior of the objective function.

4.1. **Smoothed quantile regression (LSQR).** We propose a quantile-based linear regression model that is based on the assumption of smoothness, i.e., data points whose predictor variables are similar, should have a similar response. We use this notion of smoothness as an integral part of the framework as experiments provided in Section 5 demonstrate this characteristic in the dataset used. The smoothness assumption could be described as the constraint

$$\sum_{i,j}^{n} w_{ij}(f_i - f_j)^2 < c$$

where w_{ij} is a measure of similarity between data point i and j, f the predicted value of the response variable and c is a constant.

Also, since the framework doesn't restrict the training set only to extreme data points, the smoothing component of the objective function tends to implicitly cluster data points resulting in better distinction of the response variables of an extreme valued data point and a non-extreme

valued data point. Empirical results comparing supervised quantile regression to the proposed semi-supervised model illustrate this point as shown in Section 5. The term

$$w_{ij} = \exp(-\frac{||x_i - x_j||^2}{\sigma}) \quad i, j \in [1, 2, \ldots, n]$$

is equivalent to the radial basis function and is used to capture the similarity between the predictor variables of data point i and data point j. σ is a scale parameter used to control the distance above which two data points are not considered as being highly coupled.

Assuming linear regression, $f(x_i, \beta) = x_i\beta$, the smoothing term can be reformulated as

$$\sum_{i,j}^{n} w_{ij}(f(x_i, \beta) - f(x_j, \beta))^2 = f^T \mathbf{\Delta} f = \beta^T \mathbf{\Sigma} \beta$$

where,

$$\mathbf{\Sigma} = X^T \mathbf{\Delta} X$$

$$\mathbf{\Delta} = D - W$$

and D is a diagonal matrix such that $D_{ii} = \sum_{j=1}^{n} w_{ij}$ and $W = \{w_{ij}\}|_{i,j=1}^{n}$.

Coupling smoothing with the objective function of linear qunatile regression, we end up with the following optimization problem.

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta) + \lambda \beta^T \mathbf{\Sigma} \beta$$

As can be clearly observed from the objective functions of $LSQR$, $\lambda \to 0$ results in an estimate similar to quantile linear regression while, $\lambda \to \infty$ results in the estimate of the response variable converging towards the target quantile of data. This is because a large λ would penalize any non-zero difference between f_i and f_j very harshly thereby minimizing the error by setting $f_i = \alpha, \forall i \in [1, 2, \ldots, n]$, thereby reducing the error from the second component of the equation to 0. This reduces the loss function to the following

$$f(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - \alpha), \quad \beta = (\alpha, 0, 0, \ldots, 0)^T$$

The formal proof of this is provided in the following theorem.

Theorem 1: $f(x_i, \beta) \to y_{(n\tau)}$ as $\lambda \to \infty$, $\forall i \in [1, 2, \ldots, n]$.

Proof : Let $y_{(i)}$ be the i^{th} smallest element among $y_k|_{k=1}^{n}$ and $y_{(i)} < \alpha_i <= y_{(i+1)}$. When $\lambda \to \infty$, the loss function can be rewritten in terms of α_i as follows

$$\sum_{k=1}^{i}(1 - \tau)(\alpha_i - y_{(k)}) + \sum_{k=i+1}^{n} \tau(y_{(k)} - \alpha_i) + \sum_{i,j=1}^{n} W_{ij}(\alpha_i - \alpha_i)$$

which is equivalent to minimizing

$$\tau \sum_{k=1}^{n} y_{(k)} - \sum_{k=1}^{i} y_{(k)} - (n\tau - i)\alpha_i$$

or maximizing

$$\sum_{k=1}^{i} y_{(k)} + (n\tau - i)\alpha_i = l_i$$

Therefore,

$$l_j - l_{j-1} = y_j - \alpha_{j-1} + (n\tau - j)(\alpha_{j-1} - \alpha_j)$$

Hence, $\forall j : j \leq n\tau, \quad l_j - l_{j-1} >= 0$, since $(y_j - \alpha_{j-1})$, $(n\tau - j)$ and $(\alpha_{j-1} - \alpha_j)$ are all ≥ 0. Similarly, $\forall j : j \geq n\tau$,

$$l_j - l_{j+1} = \alpha_{j+1} - y_{j+1} + (n\tau - j)(\alpha_j - \alpha_{j+1}) \geq 0$$

Hence, if $\exists i : i = n\tau$, then $\alpha = y_{(n\tau)}$. But if, $i < n\tau < (i+1)$, then α is in the interval $[y_{(i)}, y_{(i+1)}]$ □

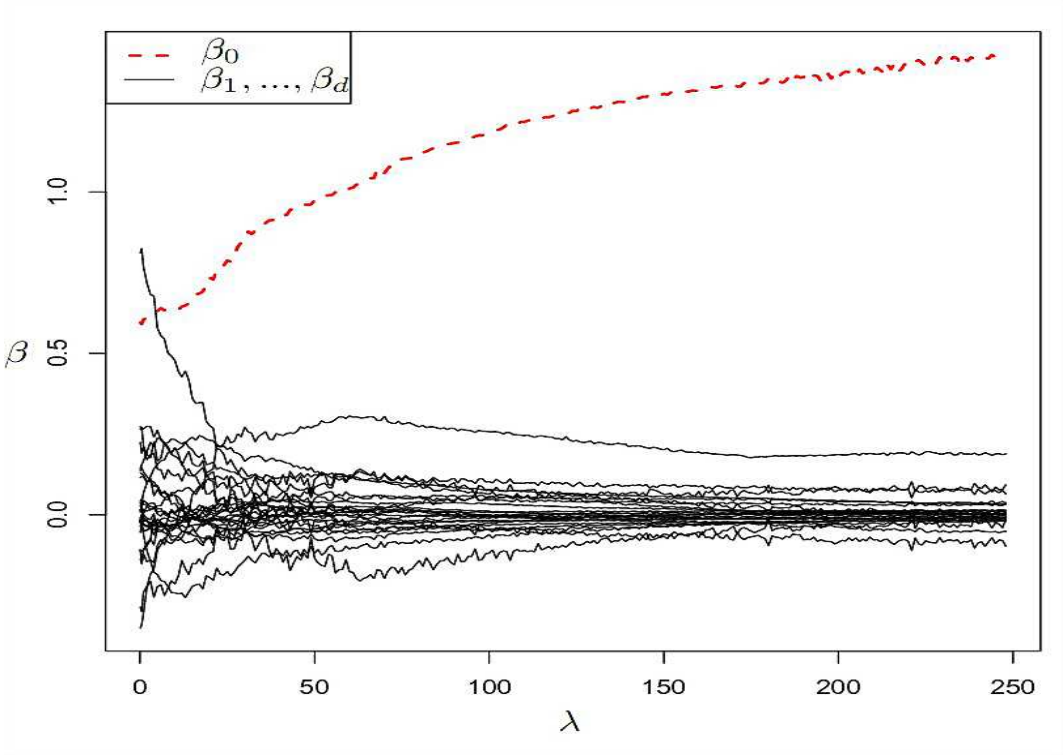

FIGURE 3. Influence of parameter λ on the regression coefficients β in LSQR.

Figure 3 is a plot that tracks the values of β for different λ values. The figure shows that the regression parameter vector $\boldsymbol{\beta}$ will converge to $(\alpha, 0, 0, \ldots, 0)^T$ as λ increases. β_0 is the regression parameter that corresponds to the column of 1's in the design matrix.

Figures 4 and 5 plots the influence of λ on the predicted values returned from LSSQR. i.e., as the value of λ increases, LSSQR shrinks the prediction range to the quantile τ. Figure 5 is a zoomed-in image, capturing the tail of Figure 4.

4.2. **Linear semi-supervised quantile regression (LSSQR).** The objective function of LSQR can be easily extended to a semi-supervised learning setting since the smoothing factor (the second term in the equation) is independent of y. Therefore, by extending the range of the indices i and j of the smoothing term to span 1 to $n + m$, the predictor variables of the unlabeled data $X_u = [x_{u1}, ..., x_{um}]^T$ can be harvested.

The objective function of the LSSQR is

$$\arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta) + \lambda \sum_{i,j}^{n+m} w_{ij}(x_i^T\beta - x_j^T\beta)^2$$

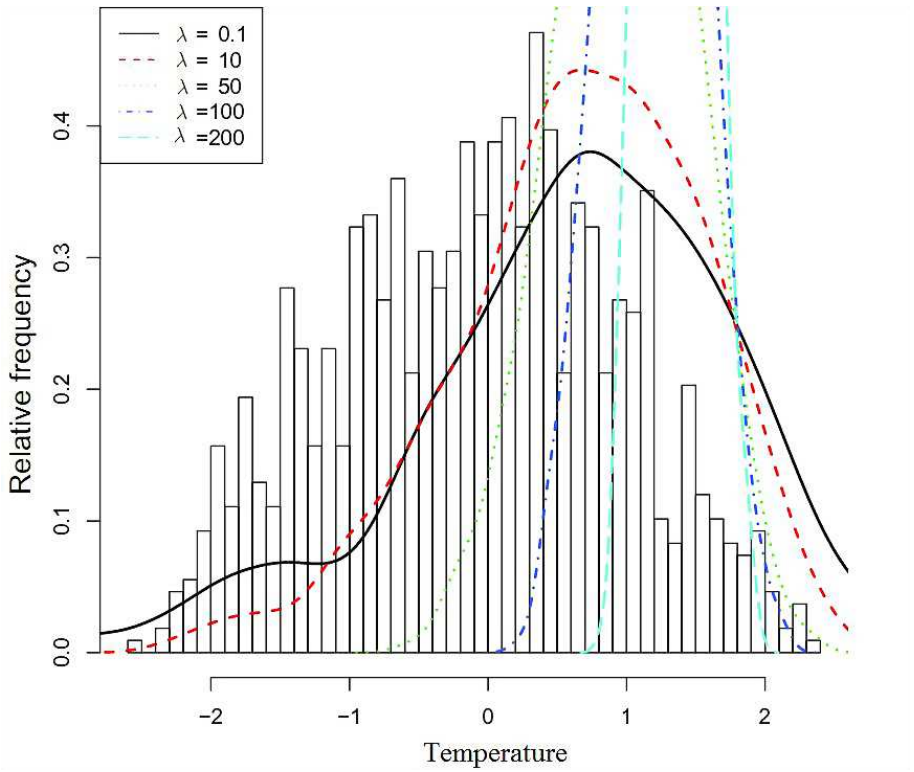

FIGURE 4. Influence of λ on the probability distribution of the predicted values obtained from LSSQR.

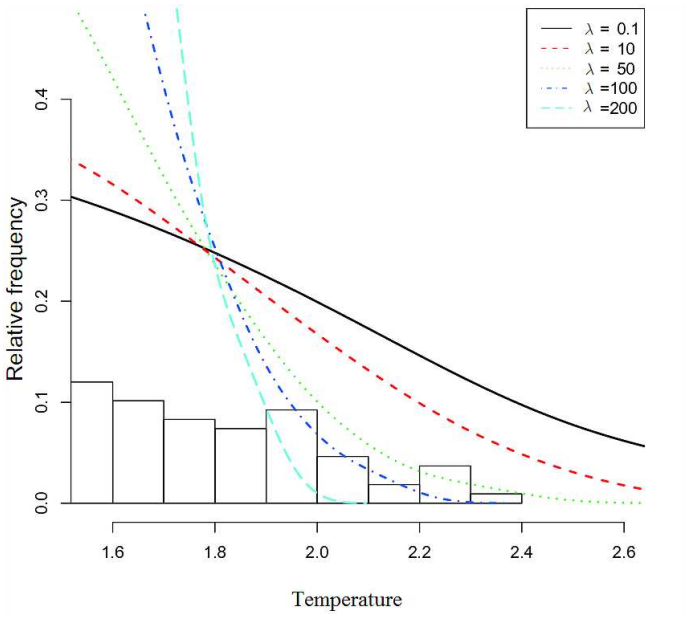

FIGURE 5. Influence of λ on the probability distribution of the predicted extreme values obtained from LSSQR.

5. Experimental Results

In this section, the climate dataset that is used for statistical downscaling is described. This is followed by the experimental setup, which address the inherent properties of the dataset, such as its periodic nature. Once the dataset is introduced, we analyze the behavior of baseline models developed using MLR, ridge regression and quantile regression and contrast them with LSQR and LSSQR. The efficacy of the models in accurately measuring the magnitude, the relative frequency and timing of forecasting a data point as an extreme event is measured.

5.1. **Data.** All the algorithms were run on climate data obtained at 37 weather stations in Canada, from the Canadian Climate Change Scenarios Network website [1]. The response variable to be regressed (downscaled) corresponds to daily temperature values measured at each weather station. The predictor variables for each of the 37 stations correspond to 26 coarse-scale climate variables derived from the NCEP re-analysis data set, which include measurements of airflow strength, sea-level pressure, wind direction, vorticity, and humidity, as shown in Table 1. The predictor variables used for training were obtained from the NCEP re-analysis data set that span a 40-year period (1961 to 2001). The time series was truncated for each weather station to exclude days for which temperature or any of the predictor values are missing.

Table 1. List of predictor variables for temperature prediction.

Predictor Variables	
500 hPa airflow strength	850 hPa airflow strength
500 hPa zonal velocity	850 hPa zonal velocity
500 hPa meridional velocity	850 hPa meridional velocity
500 hPa vorticity	850 hPa vorticity
500 hPa geopotential height	850 hPa geopotential height
500 hPa wind direction	850 hPa wind direction
500 hPa divergence	850 hPa divergence
Relative humidity at 500 hPa	Relative humidity at 850 hPa
Near surface relative humidity	Surface specific humidity
Mean sea level pressure	Surface zonal velocity
Surface airflow strength	Surface meridional velocity
Surface vorticity	Surface wind direction
Surface divergence	Mean temp at 2 m

5.2. **Experimental setup.** As is well known, temperature, which is the response variable in our experiments, has seasonal cycles. To efficiently capture the various cycles, de-seasonalization is performed prior to running the experiments. As is common practice in the field of climatology, a common approach to de-seasonalization is to split the data into 4 seasons (DJF, MAM, JJA, SON) where 'DJF' refers to the months of December-January-February in the temperature timeseries. Similarly, 'MAM' refers to March-April-May, and 'JJA' refers to June-July-August and 'SON', September-October-November. In effect, for each station, we build 4 different models, corresponding to the 4 seasons. The training size used spanned 6 years of data and the test size, 12 years. During validation, the parameter λ was selected using the score returned by RMSE for extreme data points. A data point is considered extreme if its response variable is greater than .95 percentile (Threshold-1) of the whole dataset corresponding to the station. QR was implemented using the interior point algorithm as detailed in [2]. Broyden Fletcher Goldfarb Shanno (BFGS) method was used to solve the LSQR and LSSQR optimization problem.

5.3. **Evaluation criteria.** The motivation behind the selection of the evaluation metrics was the intent to evaluate the different algorithms in terms of accuracy of the prediction of extreme values, the timing of the extreme events as well as the frequency with which a data point is predicted to be an extreme data point. The following metrics are used to capture the above evaluation criteria for the various models:

- Root Mean Square Error (RMSE), which measures the difference in magnitude between the actual and predicted values of the response variable, i.e.:
  RMSE $= \sqrt{\frac{\sum_{i=1}^{n}(y_i'-f_i')^2}{n}}$. RMSE was computed on those days that were observed to be extreme data points.
- Precision and recall of extreme events are computed to measure the timing accuracy of the prediction. F-measure, which is the harmonic mean between recall and precision values, will be used as a score that summarizes the precision and recall results.
  F-measure $= \frac{2\times Recall\times Precision}{Recall+Precision}$
- The frequency of predicting extreme data point for the various methods was measured by computing the ratio of the number of data points that were predicted to be extreme to the number of observed extreme data points.

To summarize, RMSE is used for measuring the accuracy of the predicted magnitude of the response variable, whereas F-measure can be thought of as measuring the correctness of the timing of the extreme events.

5.4. **Baseline.** We compared the performance of LSQR and LSSQR with baseline models created using multiple linear regression (MLR), ridge regression (Ridge), and quantile regression (QR). All the baselines were run for the same 37 stations and for all the 4 seasons. Also, a comparison of the performance of the proposed supervised framework (LSQR) is made with its semi-supervised counterpart (LSSQR), where LSSQR demonstrated an improved performance over LSQR for the 37 stations evaluated upon as shown in Table 2. Table 2 summarizes the tally of percentage of times LSSQR outperformed LSQR over the 4 seasons for the given 37 stations. As seen in the table, LSSQR showed an improved performance in terms of both RMSE and F-measure.

TABLE 2. The relative performance of LSSQR compared with LSQR with regard to the extreme data points.

	Win	Loss	Tie
RMSE	68.25%	31.75%	0%
F-measure	60.14%	37.16%	2.7%

5.5. **Results.** As mentioned earlier, experiments were run separately using each of the baseline approaches and LSQR and LSSQR for the 4 seasons (DJF, MAM, JJA, SON) of the year for each of the 37 stations' data. The results over all the seasons and stations are summarized in Tables 3 and 4 while the individual results of each season in Figures 6 and 8. Table 3 summarizes the relative performance of LSQR with respect to the baseline methods in terms of RMSE of extreme data points and F-measure of identification of extreme data points. During testing, a data point is considered extreme, if its response variable is greater than .95 percentile (Threshold-1) of the whole dataset corresponding to the station. For the purpose of analysis, results of using the .95 percentile of the response variable in the training set (Threshold-2) to identify extreme data points are also summarized. The fact that the results obtained by using the two different baselines is an indicator that the training data did capture the distribution of the response variable reasonably well. LSQR consistently outperformed the baselines both in terms of RMSE and F-measure. It must also be noted that LSQR did outperform MLR and Ridge in terms of recall of extreme events comprehensively across each of the 37 stations and seasons.

TABLE 3. The percentage of stations LSQR outperformed the respective baselines, with regard to the extreme data points.

		MLR	Ridge	QR
RMSE	Threshold-1	88.51%	87.84%	80.40%
	Threshold-2	89.19%	87.84%	79.05%
F-measure	Threshold-1	59.45%	60.13%	72.97%
	Threshold-2	56.08%	58.10%	79.05%

TABLE 4. The percentage of stations LSSQR outperformed the respective baselines, with regard to the extreme data points.

		MLR	Ridge	QR
RMSE	Threshold-1	87.16%	85.14%	85.13%
	Threshold-2	87.84%	86.49%	81.76%
F-measure	Threshold-1	60.13%	58.78%	75.67%
	Threshold-2	56.75%	59.45%	81.75%

Similarly, Table 4 summarizes the relative performance of LSSQR with respect to the baseline methods in terms of RMSE of extreme data points and F-measure of identification of extreme data points. Like LSQR, LSSQR consistently outperformed the baselines both in terms of RMSE and F-measure. It must be noted that LSSQR outperform MLR and Ridge in terms of recall of extreme events comprehensively across each of the 37 stations and seasons.

Figure 6 gives a breakdown of the performance of the LSSQR over each of the 4 seasons of the 37 stations using Threshold-1 for the purpose of marking a data point as extreme. The figure is a bar chart of percentage of stations that LSSQR outperformed MLR, ridge regression and QR in prediction accuracy for only extreme data points in the test set. RMSE was used to compute the accuracy of each model in predicting extreme value data points, at the 37 stations. As seen in the plot, LSSQR outperforms MLR, ridge regression and QR in each of the four seasons across the 37 stations.

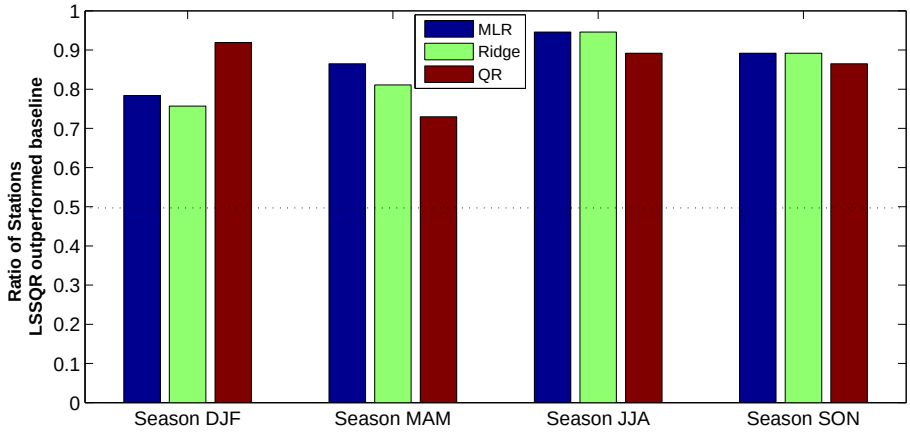

FIGURE 6. Ratio of stations LSSQR outperforming baseline in terms of RMSE of extreme data points.

Figure 7 shows a graph that depicts the percentage of stations LSSQR outperformed MLR, ridge regression and QR in terms of identifying extreme data points over 37 stations. Again, LSSQR

comprehensively outperforms MLR and ridge regression over all the 37 stations and 4 seasons. But as expected, QR outperforms LSSQR in terms of recall performance for each of the 4 seasons due to the overestimating nature of QR, which consequently resulted in poor precision and which is reflected in its F-measure score. At an average, quantile regression, predicted a datapoint to be an extreme point more than twice as frequently as the actual frequency of observed extreme data points. In fact, QR lost out to LSSQR in 91% of 37 stations across 4 seasons in terms of precision of identifying extreme data points.

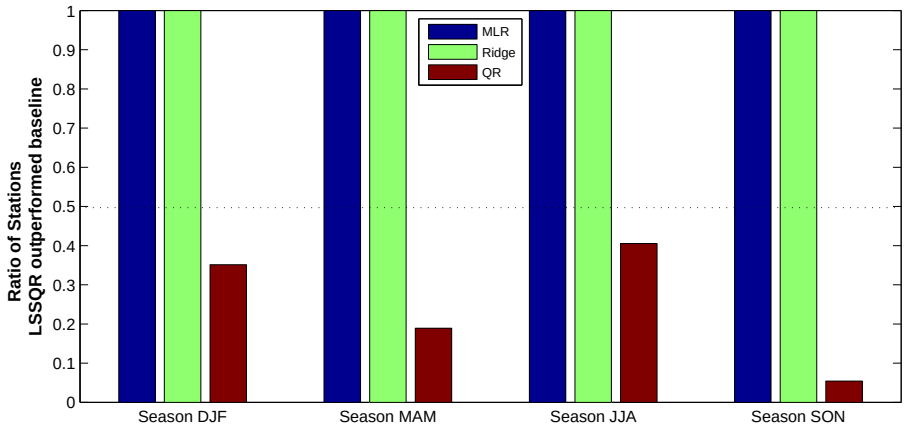

FIGURE 7. Ratio of stations LSSQR outperforming baseline in terms of recall of extreme data points.

Figure 8 shows a graph that depicts the percentage of stations where LSSQR outperformed MLR, ridge regression and QR in prediction accuracy based on F-measure of the identifying extreme data points over 37 stations. Again, LSSQR outperforms MLR, ridge regression and QR for all the 4 seasons.

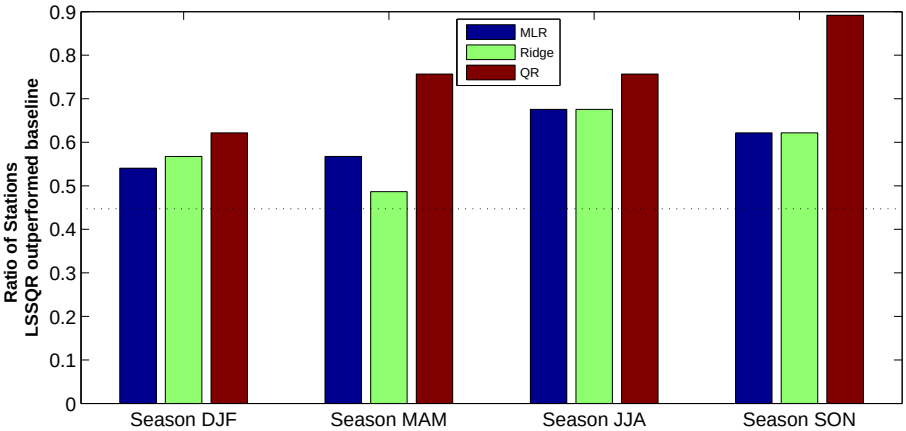

FIGURE 8. Ratio of stations LSSQR outperformin baseline in terms of F-measure of extreme data points.

The performance improvement obtained by LSSQR in terms of predicting the extreme values can be easily visualized in Figure 9. Figure 9 is a plot comparing the predicted response variable of the various methods. The plot is restricted to only extreme data points for a station. As expected, the predicted value of the response variable using multiple linear regression is often underestimating the observed temperature, while quantile regression regularly overestimates the prediction of temperature and LSSQR lies in between MLR and QR and closer to the observed temperature.

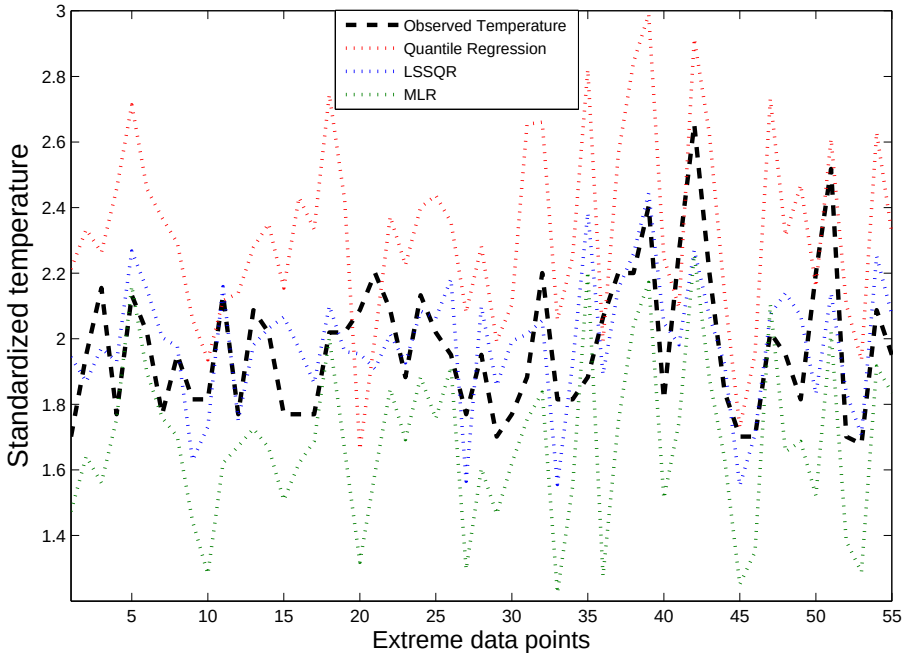

FIGURE 9. Prediction performance of extreme data points using MLR, Ridge, QR, LSSQR.

6. Conclusions

This paper presents a semi-supervised framework (LSSQR) for recalling and accurately predicting values for extreme data points. The proposed approach was applied to real world climate data spanning 37 stations and was compared with MLR, ridge regression and quantile regression in terms of the effectiveness the model demonstrated in identifying and predicting extreme temperatures for the given stations. For future work, we will explore a non-linear variant of the smoothed quantile regression framework. We will also explore a semi-supervised variant of the non-linear smoothed quantile regression model.

References

[1] Canadian Climate Change Scenarios Network, Environment Canada. http://www.ccsn.ca/
[2] R. Koenker. Quantile Regression Software. http://www.econ.uiuc.edu/ roger/research/rq/rq.html
[3] S. Ancelet, M.-P. Etienne, H. Benot, and E. Parent. Modelling spatial zero-inflated continuous data with an exponentially compound poisson process. *Environmental and Ecological Statistics*, DOI:10.1007/s10651-009-0111-6, April 2009.
[4] S. Barry and A. H. Welsh. Generalized additive modelling and zero inflated count data. *Ecological Modelling*, 157(2-3):179–188, November 2002.

[5] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proc. of the 18th Int'l Conf. on Machine Learning*, pages 19–26, 2001.
[6] J. Bjornar Bremnes, Probabilistic Forecasts of Precipitation in Terms of Quantiles Using NWP Model Output. In *Monthly Weather Review*, pages 338–347, 2004
[7] N. Meinshausen. Quantile Regression Forests In *Journal Machine Learning*, pages 983-999, 2006
[8] L. Feudale. Large scale extreme events in surface temperature during 1950–2003: an observational and modeling study In *Ph.D. Dissertation. George Mason University*
[9] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. of the Workshop on Computational Learning Theory*, pages 92–100, 1998.
[10] D. Bohning, E. Dierz, and P. Schlattmann. Zero-inflated count models and their applications in public health and social science. In J. Rost and R. Langeheine, editors, *Applications of Latent Trait and Latent Class Models in the Social Sciences*. Waxman Publishing Co, 1997.
[11] Y.-A. L. Borgne, S. Santini, and G. Bontempi. Adaptive model selection for time series prediction in wireless sensor networks. *Signal Process*, 87(12):3010–3020, 2007.
[12] U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel. Efficient co-regularised least squares regression. In *Proc. of the 23rd Int'l Conf. on Machine learning*, pages 137–144, 2006.
[13] S. Charles, B. Bates, I. Smith, and J. Hughes. Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. In *Hydrological Processes*, pages 1373–1394, 2004.
[14] Z. Abraham and P.-N. Tan. An Integrated Framework for Simultaneous Classification and Regression of Time-Series Data. In *Proc of the ACM SIGKDD Int'l Conf on Data Mining*, Colorado, OH, 2010.
[15] H. Cheng and P.-N. Tan. Semi-supervised learning with data calibration for long-term time series forecasting. In *Proc of the ACM SIGKDD Int'l Conf on Data Mining*, Las Vegas, NV, 2008.
[16] I. Cohen, N. Sebe, F. G. Cozman, M. C. Cirelo, and T. S. Huang. Semi-supervised learning of classifiers: Theory and algorithms for bayesian network classifiers and applications to human-computer interaction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(12):1553–1566, Dec 2004.
[17] C. Cortes and M. Mohri. On transductive regression. In *Advances in Neural Information Processing Systems*, 2006.
[18] F. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. In *Proc. of the 15th Int'l Florida Artificial Intelligence Society Conference*, pages 327–331, 2002.
[19] F. Cozman, I. Cohen, and M. Cirelo. Semi-supervised learning of mixture models. In *Proc of the 20th Int'l Conf. on Machine Learning*, 2003.
[20] W. Enke and A. Spekat. Downscaling climate model outputs into local and regional weather elements by classification and regression. In *Climate Research 8*, pages 195–207, 1997.
[21] D. Erdman, L. Jackson, and A. Sinko. Zero-inflated poisson and zero-inflated negative binomial models using the countreg procedure. In *SAS Global Forum 2008*, pages 1–11, 2008.
[22] P. Friederichs and A. Hense Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression. In *Monthly Weather Review*, pages 2365–2378, 2007
[23] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama. Multiple-regression hidden markov model. In *Proc. of IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, 2001.
[24] C. Giles, S. Lawrence, and A. Tsoi. Noisy time series prediction using a recurrent neural network and grammatical inference. *Machine Learning, 44(1-2)*, pages 161–183, 2001.
[25] W. Hong, P. Pai, S. Yang, and R. Theng. Highway traffic forecasting by support vector regression model with tabu search algorithms. In *Proc. of Int'l Joint Conf. on Neural Networks*, pages 1617–1621, 2006.
[26] T. Joachims. Transductive inference for text classification using support vector machines. In *Proc. of the 16th Int'l Conf. on Machine Learning*, pages 200–209, Bled, SL, 1999.
[27] R. Koenker and K. Hallock. Quantile Regression. *Journal of Economic PerspectivesVolume 15, Number 4*, pages 143-156, 2001.
[28] B. Kedem and K. Fokianos. Regression models for time series analysis. *Wiley-Interscience ISBN: 0-471-36355*, 2002.
[29] E.C. Mannshardt-Shamseldin, R.L. Smith, S.R. Sain, L.D. Mearns and D. Cooley Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data. In *Annals of Applied Statistics*, pages 484–502, 2010.
[30] T. Martin, B. Wintle, J. Rhodes, P. Kuhnert, S. Field, S. Low-Choy, A. Tyre, and H. Possingham. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, 8:1235–1246, 2005.
[31] N. Meinshause. Quantile Regression Forests. *Journal of Machine Learning Research 7*, 7:9839-99, 2006.
[32] C. Monteleoni, G. Schmidt AND S. Saroha Tracking Climate Models. *NASA Conference on Intelligent Data Understanding (CIDU)*, 2010.
[33] A. Ober-Sundermeier and H. Zackor. Prediction of congestion due to road works on freeways. In *Proc. of IEEE Intelligent Transportation Systems*, pages 240–244, 2001.

[34] A. Smola and B. Scholkopf. A tutorial on support vector regression. In *Statistics and Computing*, pages 199–222(24). Spring, 2004.

[35] I. Takeuchi, Q.V. Le, T. Sears and A.J. Smola. Nonparametric Quantile Regressio. In *Journal of Machine Learning Research Nonparamteric Quantile Estimation*,2005.

[36] Q. Tian, J. Yu, Q. Xue, and N. Sebe. A new analysis of the value of unlabeled data in semi-supervised learning for image retrieval. In *Proc. of IEEE Int'l Conf. on Multimedia and Expo.*, pages 1019– 1022, 2004.

[37] E. Towler, B. Rajagopalan, E. Gilleland, R.S. Summers, D. Yates, and R.W. Katz Modeling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory. In *Water Resources Research*, VOL. 46, W11504, 2010

[38] L. Wei and E. J. Keogh. Semi-supervised time series classification. In *Proc of ACM SIGKDD Int'l Conf on Data Mining*, pages 748–753, Philadelphia, PA, August 2006.

[39] A. H. Welsh, R. Cunningham, C. Donnelly, and D. B. Lindenmayer. Modelling the abundance of rare species: statistical models for counts with extra zeros. In *Ecological Modelling*. Elsevier, Amsterdam, PAYS-BAS (1975) (Revue), 1996.

[40] R. Wilby, S. Charles, E. Zorita, B. Timbal, P. Whetton, and L. Mearns. Guidelines for use of climate scenarios developed from statistical downscaling methods. Available from the DDC of IPCC TGCIA, 2004.

[41] C.-C. Wong, M.-C. Chan, and C.-C. Lam. Financial time series forecasting by neural network using conjugate gradient learning algorithm and multiple linear regression weight initialization. Technical Report 61, Society for Computational Economics, Jul 2000.

[42] L. Youjuan,L. Yufeng, and Ji Z HU Quantile Regression in Reproducing Kernel Hilbert Spaces In *American Statistical Association Vol. 102, No. 477, Theory and Methods*, 2007

[43] T. Zhang. The value of unlabeled data for classification problems. In *Proc of the Int'l Conf. on Machine Learning*, 2000.

[44] Z. Zhou and M. Li. Semi-supervised regression with co-training. In *Proc. of Int'l Joint Conf. on Artificial Intelligence*, 2005.

[45] X. Zhu. Semi-supervised learning literature survey. In *Technical Report,Computer Sciences, University of Wisconsin-Madison*, 2005.

[46] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. of the 20th Int'l Conf. on Machine Learning*, volume 20, 2003.

[47] X. Zhu and A. Goldberg. Kernel regression with order preferences. In *Association for the Advancement of Artificial Intelligence*, page 681, 2007.

[48] C. Dorland, R.S.J. Tol and J.P. Palutikof. Vulnerability of the Netherlands and Northwest Europe to storm damage under climate change. In *Climatic Change*, pages 513-535, 1999.

[49] Y. Hundecha, A. St-Hilaire, T.B.M.J. Ouarda, S. El Adlouni, and P. Gachon. A nonstationary extreme value analysis for the assessment of changes in extreme annual wind speed over the Gulf of St. Lawrence, Canada. In *Journal of Applied Meteorology and Climatology*, pages 2745-2759, 2008.

[50] M.J. Booij, Extreme daily precipitation in Western Europe with climate change at appropriate spatial scales. In *International Journal of Climatology*, 2002.

[51] N.B. Bernier, K.R. Thompson, J. Ou, and H. Ritchie. Mapping the return periods of extreme sea levels: Allowing for short sea level records, seasonality, and climate change. In *Global and Planetary Change*, pages 139-150, 2007.

[52] R.T. Clarke. Estimating trends in data from the Weibull and a generalized extreme value distribution. In *Water Resources Research*, 2002.